(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
28 December 2000 (28.12.2000)

PCT

(10) International Publication Number
WO 00/79426 A1

(51) International Patent Classification⁷: G06F 17/21

(21) International Application Number: PCT/US00/40238

(22) International Filing Date: 19 June 2000 (19.06.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/139,930 18 June 1999 (18.06.1999) US

(71) Applicant *(for all designated States except US)*: **THE TRUSTEES OF COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK** [US/US]; 116th Street and Broadway, New York, NY 10027 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **KLAVANS, Judith, L.** [US/US]; 40 South Drive, Hastings-on Hudson, NY 10706 (US). **ESKIN, Eleazar** [—/US]; Columbia University, Shapiro Room 722, 116th Street and Broadway, New York, NY 10027 (US). **HATZIVASSILOGLOU, Vasileios** [—/US]; Columbia University, Shapiro Room 724, 116th Street and Broadway, New York, NY 10027 (US).

(74) Agents: TANG, Henry et al.; Baker Botts LLP, 30 Rockefeller Plaza, New York, NY 10112-0228 (US).

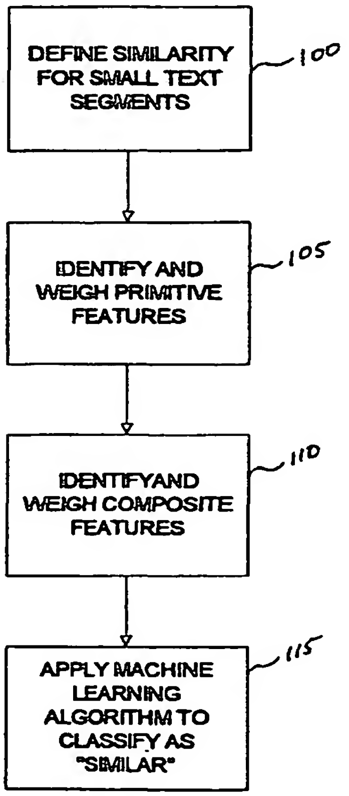
(81) Designated States *(national)*: JP, US.

(84) Designated States *(regional)*: European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published:
— *With international search report.*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: SYSTEM AND METHOD FOR DETECTING TEXT SIMILARITY OVER SHORT PASSAGES


DEFINE SIMILARITY FOR SMALL TEXT SEGMENTS
100
IDENTIFY AND WEIGH PRIMITIVE FEATURES
105
IDENTIFYAND WEIGH COMPOSITE FEATURES
110
APPLY MACHINE LEARNING ALGORITHM TO CLASSIFY AS "SIMILAR"
115

(57) Abstract: A system and method are provided for determining similarity in short text segments. The method provides a definition of similarity which is appropriate for the small text setting (100). Small text segments are compared to determine if there exist common primitive features, such as words, noun phrases, synonyms, verbs with a common semantic class, proper nouns and the like (105). From the primitive features identified, the small text segments are evaluated to determine whether composite features are present (110). Composite features are defined as predetermined relationships between primitive features. The common primitive features and composite features are applied as inputs to an appropriate machine learning algorithm which is trained to ascertain a similarity measure based on the primitive and composite features common to the text segments (115).

6/pyts

SYSTEM AND METHOD FOR DETECTING TEXT SIMILARITY OVER SHORT PASSAGES

FIELD OF THE INVENTION

The present invention relates generally to natural language processing and more particularly relates to a system and method for determining the similarity of text in short passages.

BACKGROUND OF THE INVENTION

With the growing volume of textual information, such as newspaper articles, magazines, Internet articles, and the like, there is a growing need to automatically cluster and/or classify such documents and determine whether groups of documents express similarities or not. For the most part, research in this area has focused on detecting similarity between documents and large segments of text or between a short query phrase and one or more documents.

While effective techniques have been developed for document clustering and classification which depend on inter-document similarity measures, these techniques generally rely only on shared words, or occasionally on collocation of words. Such techniques are applicable when large units of text, such as full documents, are compared. In this case, there is generally sufficient overlap to detect similarity in the documents and/or document segments. However, when the units of text are small, for example a paragraph or abstract, such simple surface matching of words and phrases is far more prone to error. In the case of small text units, the sample size is reduced and the number of potential matches is reduced accordingly. Thus, there remains a need for improved techniques for detecting similarities between small text units.

A further problem with known techniques for detecting similarity is that the conventional notions of similarity which are applicable to large text samples, such as documents and large text segments, do not provide sufficient measures of similarity for measuring similarity in small text segments. Standard notions of similarity generally involve the creation of a vector or profile of characteristics of a text fragment and determine a conceptual distance between vectors on the basis of frequencies. Features typically include stemmed words, although multi-word units and collocations also have been used. Typological characteristics, such as thesaural features, have also been used to calculate features. The difference between vectors for one text unit (usually a query) and another text unit (usually a document) then determines closeness or similarity of the text units.

In some cases, the text units are represented as vectors of sparse n-grams of word occurrences and learning is applied over those vectors. Though effective in the context of large document comparisons, a more fine-grained distinction for similarity measures is required to properly characterize the similarity of two small text segments.

SUMMARY OF THE INVENTION

It is an object of the present invention to provide systems and methods for detecting similarity between two or more small text segments.

A method for determining similarity in short text segments in accordance with the present invention includes the steps of determining common primitive features in the text segments, determining common composite features in the text segments and then calculating a similarity measure based upon the primitive and composite features. The primitive features can be selected from the group including common single words, common noun phrases, synonyms, common semantic classes of verbs, and common proper nouns. The composite features, which represent relationships between and among the primitive features, can be selected from the group including primitive feature order restrictions, primitive feature distance restrictions, and primitive type restrictions.

Preferably, the step of determining common primitive features can include the further steps of identifying common primitive features, assigning a value to the primitive features, and normalizing the feature values. Normalizing the values can include normalizing for text segment length and normalizing for the frequency of primitive feature occurrence. Similarly, determining composite features generally includes identifying the composite features, assigning a value to the composite features, and normalizing the feature values. Again, normalization of the feature values can include normalizing for text segment length and normalizing for the frequency of feature occurrence.

BRIEF DESCRIPTION OF THE DRAWING

Further objects, features and advantages of the invention will become apparent from the following detailed description taken in conjunction with the accompanying figures showing illustrative embodiments of the invention, in which

Figure 1 is a flow chart illustrating an overview of a present method for comparing small text segments;

Figure 2 is a flow chart illustrating the step of defining similarity for small text segments in accordance with the present methods;

Figure 3 is a flow chart illustrating the process of computing primitive features for use in detecting similarity in small text segments;

Figure 4 is a flow chart illustrating the process of calculating composite features for use in detecting similarity of small text segments in accordance with the present methods;

Figure 5 is a block diagram of a software system topology for determining similarity in small text segments in accordance with the present methods;

Figure 6 is an illustration of exemplary short text segments;

Figure 7 is a diagram illustrating a composite feature match between two of the short text segments provided in Figure 6 using a "same order" rule;

Figure 8 is a diagram illustrating a composite feature match between two of the short text segments provided in Figure 6 using a "within distance" rule; and

Figure 9 is a diagram illustrating a composite feature match between two of the short text segments provided in Figure 6 using a "primitive type" rule.

Throughout the figures, the same reference numerals and characters, unless otherwise stated, are used to denote like features, elements, components or portions of the illustrated embodiments. Moreover, while the subject invention will now be described in detail with reference to the figures, it is done so in connection with the illustrative embodiments. It is intended that changes and modifications can be made to the described embodiments without departing from the true scope and spirit of the subject invention as defined by the appended claims.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

Figure 1 is a flow chart illustrating an overview of the process used in the present invention for detecting similarity in small text segments. As previously noted, a problem in the prior art is that the definition of similarity commonly used for large text segments, such as documents, is not sufficiently refined to provide an adequate measure of similarity when comparing small text segments. Generally, small text segments refer to sentences, phrases and short paragraphs.

Referring to Figure 1, in step 100 a definition of similarity for small text segments is provided. From this definition, the method proceeds to identify primitive features of the small text segments and determine feature values for the primitive features (step 105). Primitive features are those which generally compare simple parts of speech and text, such as single words, word categories, or phrases such as noun phrases, synonyms, verb class and proper nouns. In addition to primitive features, the process can identify composite features of the short-text segments and determine composite feature values (step 110). Composite features are those which compare relationships among two or more primitive features. Once primitive features and composite features have been identified and given an appropriate value, a machine learning algorithm is applied to classify small text segments as similar or not similar (step 115).

Figure 2 is a flow chart which illustrates the process of establishing an appropriate definition of similarity for small text segments. In general, two text units can be considered as similar if they share the same focus on a common concept, actor, object or action. In addition, the common actor or object definition must perform or be subjected to the same action or be the subject of the same description. This is exemplified in the flow chart of Figure 2, where two small text segments are selected from a body of text and are analyzed. If the two text segments relate to a common concept (step 205), then further analysis is performed to see if the common concept relates to the same action (step 210) or relates to the same description (step 215).

Similar tests are performed to determine if the two text segments relate to a common actor (step 220) or to a common object (step 225). If there is no common concept, actor or object, the text segments are considered not similar (step 235). Similarly, for those text segments which do refer or relate to a common concept, actor or object, those segments will still be found not similar unless they also relate to a common action or involve the same description. Thus, for short text segments to be similar, they must contain a common concept, actor, or object which is also the subject of a common action or description. The comparisons in steps 205, 220 and 225 can be the basis for primitive features 240. Those relationships between primitive features which are identified in steps 210, 215 can be referred to as composite features 245.

While Figure 2 is illustrated as a sequential process, it represents a decision tree involved in a definition of similarity of two short text segments as applied in the present invention which can also be performed in a largely parallel manner. For example, decisions 205, 220 and 225 can be performed concurrently as can decisions 210 and 215. Using this definition of similarity for small text segments, a feature-based process can be employed which compares primitive and composite features of short text segments to determine if the definition is satisfied for two or more given input text segments.

Figure 3 is a flow chart which illustrates a method for extracting and scaling primitive features in accordance with the present invention. The text segments are compared for a level of commonality, including determining whether there is a common single word (step 305), a common noun phrase (step 310), whether two words in the phrases are synonyms (step 315), whether the phrases include verbs having a common semantic class (step 320), and whether a common proper noun can be found in the two phrases (step 325). If none of these conditions are satisfied for the applied small text segments, there is no primitive feature common to these two text segments (step 327). When a primitive feature has been identified, e.g., one of the conditions in steps 305 through 325 is satisfied, a feature value is assigned to that primitive feature. Preferably, the values which are assigned to the features are determined by a machine learning algorithm, such as RIPPER, which is trained using a suitable training corpus. RIPPER is a widely-used and effective rule induction system which is available from AT&T Laboratories and is described by Cohen in "Learning Trees and Rules with Set-Valued Features, Proceedings of the Fourteenth National Conference on Artificial Intelligence, American Association on Artificial Intelligence, 1996, which is incorporated by reference. It has been found that a subset of a corpus of 264 paragraphs which have been manually tagged by human readers as similar or not similar can be used to establish a feature rule set for RIPPER which is then suitable for assigning values to the features identified in the text segments. The particular training corpus and learned rule set will generally vary depending on the desired application. The values assigned will vary based on properties of the machine learning algorithm and training corpus. After feature values are assigned in step 330, these values can be normalized based on text length (step 335) and/or noted frequency of occurrence (step 340). Though normalization is optional, it is a desirable step to provide uniform and accurate results across varying types of text and length of text segments.

Primitive features provide a baseline indication of similarity. To further refine the notion of similarity in small text segments, relationships among primitive features, referred to as composite features, can also be evaluated. Referring to Figure 4, a method of evaluating composite features is illustrated. Composite features are those features which identify relationships among primitive feature pairs. Generally, composite features are defined by placing different forms of restrictions on participating primitive feature pairs. Referring to Figure 4, the primitive features identified in each of the small text segments are applied to a test layer 400 where various feature relationships are evaluated. The relationships illustrated in test layer 400 are exemplary in nature and are not intended to illustrate an exhaustive list of possible relationships. It will be appreciated that an large number of relationships between and among primitive features can be used to establish composite features.

For example, one type of feature relationship for composite features can be that the primitives occur in the same order in each of the text samples (step 405). This is illustrated by example in Figure 7. Figure 6 provides three short text segments to be compared. Figure 7 illustrates a match according to the "same order" composite feature rule. In Figures 7-9, primitive features are identified by shading and the relationships which form the composite features are illustrated by connecting lines. In the case illustrated in Figure 7 the primitive features {two, contact} appear in the same order in text segments Figure 6 (a) and 6 (b) from Figure 6.

Another possible relationship is that two pairs of primitive elements are required to occur within a certain distance in both text segments. The maximum distance between the primitive elements which would satisfy the relationship can be a variable or a predetermined constant (step 410). Referring to Figure 8, an example of a positive match for the "within distance" composite feature rule is provided, given that the distance, *n*, is set to a value less than three. In Figure 8, although the primitive features {contact, lost} do not appear in the same order, they occur within *n* words of each other ($n<3$ in this case).

Yet another exemplary relationship can be that the two text segments include the same primitive feature types. For example, one primitive feature can be restricted to a simplex noun phrase while the other to a verb. In such a case, two noun phrases, one from each text unit, must match according to the rule for matching simplex noun phrases and two verbs must match according to the applied rules of verb primitives (e.g., sharing the same semantic class). This is illustrated in Figure 9 where the primitive feature "An OH-58 helicopter" is deemed a simplex noun phrase match with "the helicopter" and both phrases include a common verb, "lost".

By matching primitive feature types, a simple grammatical relationship is determined in the text segments. Returning to Figure 4. for each condition that is satisfied in test layer 400, feature values are assigned to those composite features identified (step 420). The feature values are assigned by a machine learning algorithm, such as RIPPER, which has been trained on a suitable training corpus. As in the case of primitive features, optionally, the feature values assigned to the composite feature can be normalized for text length and relative occurrence of the primitive feature or composite feature (steps 425, 430, respectively). Once both primitive features and composite features of the small text segments have been identified, a machine learning algorithm is applied to determine a similarity value between the text segments (step 435). The machine learning algorithm can perform a rule-based analysis to determine similarity. Alternatively, a simpler algorithm can be used to determine similarity by comparing the total feature value of the text segments being compared to a predetermined threshold value.

Figure 5 is a block diagram of an exemplary software system for conducting the method described in connection with Figures 1-4. The system is generally implemented in software for a general purpose computer, such as a personal computer or work station. The system includes a main processing section 500. One or more interface modules 510 are included for receiving text input for the text segments to be compared and for providing the text segments to the main processing section 500. The text input can be provided by a number of sources, including but not limited to, computer readable memory, hard disks, optical disks, network databases, on-line sources, manual keyed input and the like. Based on the desired text source and input mechanism, one skilled in the art can provide appropriate text input interface module 510 hardware and software.

The main processing section 500 is also operatively coupled to a training corpus 515, which is generally stored in computer readable storage media. The main processing section 500 is generally programmed in a structured manner which calls various subprograms, library routines, and the like to perform the various functions described in accordance with Figures 1-4. The main processing section 500 can invoke the various subroutines sequentially (serial) or in a parallel, or batched, processing mode. The received text is generally passed to a preprocessing routine 520. The preprocessing routine cleans up the received text, such as by removing control characters from the text. The preprocessing routine also performs part-of-speech (POS) tagging, using known techniques, such as are available in the ALEMBIC tool set, described by Aberdeen et al. in "MITRE: Description of the Alembic System as used for MUC-6," Proceedings of the Sixth Message Understanding Conference, 1995, which is hereby incorporated by reference. ALEMBIC provides a set of data and language processing tools which identify the various parts of speech present in the small text segments.

Following text preprocessing, control is returned to the main processing section 500 which then preferably invokes a noun phrase comparison subroutine 525, such as LinkIt, to perform noun phrase comparison of step 310. LinkIt can be employed to determine whether a common noun phrase is present in the applied text segments and for identifying simplex noun phrases and matching those that share the same noun head. The LinkIt tool is described by N. Wacholder in "Simplex NPs Clustered by Head: A Method for Identifying Significant Topics in a Document", Proceedings of the Workshop on the Computational Treatment of Nominals, October 1998, which is hereby incorporated by reference in its entirety.

To determine if two segments include common proper nouns as required in step 325, the noun comparison algorithm can also be used to match those nouns identified using the ALEMBIC toolset using various predetermined matching criteria. Variations on proper noun matching can include restricting the proper noun type to a person, place or organization. Such subcategories can also be extracted using ALEMBIC's named entity finder.

Following noun phrase identification and matching, other routines for detecting primitive features can be employed. For example, to perform step 305 and determine whether common single word primitive features exist between two text segments, a word co-occurrence detection sub-routine 540 can be called by the main program 500. Variations of the word co-occurrence operation can restrict matching to cases where the parts of speech of the words also match, or relax the comparison to cases where only the word stems of the two words are identical.

Similarly, to determine if two text segments include words which are synonyms, a synonym detection algorithm 530 can be called by the main processing routine 500. In this regard, a lexical database such as WordNet®, as described by G. Miller in "WordNet, An On-Line Lexical Database," International Journal of Lexicography, Vol. 3, No. 4 (1990), can be employed. WordNet provides sense information and places words in sets of synonyms (synsets). Words that appear in the same synset are generally considered matches. Variations on this feature can be used to restrict the words being compared to a specific part-of-speech class.

To determine if two verbs present in the short text segments are of the same semantic class as set forth in step 320, a verb classifier and comparator algorithm 535 can be operatively coupled to the main processing section 500 and called by the main program. Semantic classes for verbs have been found to be useful for determining document types and text similarity. This is discussed, for example, in "The Role of Verbs in Document Analysis" by J. Klavans et al., Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics, 1998, which is hereby incorporated by reference in its entirety. For those verbs which are found to have a common semantic class, e.g., communication, motion, agreement, argument, etc., those verbs are considered to match.

The program operating in main processing section 500 can also provide algorithms to normalize feature values for text lengths and relative occurrence of the primitive. To normalize feature values for text length, as set forth in step 335, each feature value can be normalized by the size of the textual segments in the pair. For example, for a pair of textual segments A and B, the feature values assigned are divided by a normalization value, N:

$$N = \sqrt{Length(A)\ x\ Length(B)} \qquad (1)$$

This operation removes any potential bias in favor of longer text segments. It is noted that the units involved in the lengths of A and the lengths of B are generally measured by a word count.

Normalization of feature values can also be based on the relative frequency of occurrence of each primitive feature. Such normalization is motivated by the general observation that infrequently matching primitive elements are likely to have a higher impact on similarity than primitives which match more frequently. Such normalization is similar to the document frequency component of the commonly employed TF*IDF calculation. In this case, each primitive feature is associated with a value which is equal to the number of textual units in which the primitive appeared in the corpus. For a primitive element which compares single words, this is the number of text segments which contain that word in the corpus; for a noun phrase, this is the number of textual units that contain noun phrases that share the same head; and similarly for other primitive types. We multiply each feature's value by:

$$Log(\frac{T}{N}) \qquad (2)$$

where T is a number of textual segments and N is the number of textual segments containing the primitive. It is noted that since normalization for text length and frequency of occurrence are both optional operations, when these two normalization techniques are selectively applied, there are up to four variations of normalizations for each primitive feature. Of course, other normalization techniques may be added to, or substituted for, the two methods discussed herein.

The program in main processing section 500 generally employs a machine learning algorithm 545 to determine whether the text units match overall. A suitable machine learning algorithm is RIPPER, as disclosed by Cohen in "Learning Trees and Rules with Set-Valued Features, Proceedings of the Fourteenth National Conference on Artificial Intelligence, American Association on Artificial Intelligence, 1996, which is incorporated by reference. RIPPER is a widely-used and effective rule induction system. This RIPPER algorithm is trained over a corpus of manually marked pairs of text units continued in the training corpus 515. A suitable corpus was constructed using a subset of the Topic Detection and Tracking (TDT) corpus developed by NIST and DARPA. The TDT corpus in a collection of over 16,000 news articles from Reuters and CNN where many of the articles have been manually grouped into 25 categories each of which correspond to a single event. The selected corpus was formed using the Reuters' articles in five of the twenty five categories from randomly selected days. The resulting training corpus 515 contained 30 related articles. The 30 articles provided 264 paragraphs which were selected as the small text segments and resulted in 10,345 comparisons between segments.

Although use of a machine learning algorithm is preferred, other algorithms can also be used. For example, an algorithm can add the total value of composite features found in the text segments and compare this value against a similarity threshold. Similarly, although it is preferred to determine feature values based on the use of a machine learning algorithm, feature values can be predetermined based on human experience through the use of a look-up table. Alternatively, all features can be given a binary value and the similarity comparison can be determined based on a simple accumulated count of detected primary and composite features.

The present methods, while evaluated on a corpus of English language documents, are not language specific and are generally applicable to any language. Of course, the individual subroutines may require some alteration to accommodate the varied constructions found in different languages.

The methods for determining similarity in small text segments described herein form an important component in larger systems, such as document archiving systems and multi-document summarization systems.

Although the present invention has been described in connection with specific exemplary embodiments, it should be understood that various changes, substitutions and alterations can be made to the disclosed embodiments without departing from the spirit and scope of the invention as set forth in the appended claims.

CLAIMS

1. A method for determining similarity in short text segments comprising:

determining common primitive features in the text segments;

determining common composite features in the text segments; and

calculating a similarity measure based upon said primitive and composite features.

2. The method for determining similarity as defined by claim 1, wherein said primitive features are selected from the group including common single word, common noun phrase, synonyms, common semantic class of verbs, and common proper nouns.

3. The method for determining similarity as defined by claim 1, wherein said composite features are selected from the group including primitive feature order restrictions, primitive distance restrictions, and primitive type restrictions.

4. The method for determining similarity as defined by claim 1, wherein said step of determining common primitive features includes:

identifying common primitive features;

assigning a value to said primitive features; and

normalizing said value.

5. The method for determining similarity as defined by claim 4, wherein said step of normalizing includes at least one of normalizing for text segment length and normalizing for frequency of primitive occurrence.

6. The method for determining similarity as defined by claim 1, wherein said step of determining common composite features includes:

identifying common primitive features;

assigning a value to said primitive features; and

normalizing said value.

7. The method for determining similarity as defined by claim 6, wherein said step of normalizing includes at least one of normalizing for text segment length and normalizing for frequency of primitive occurrence.

---

8. A system for determining similarity in short text segments comprising:

an interface circuit for receiving text segments for comparison;

a main processing section, the main processing section being operatively couple to the interface circuit and operating under the control of a computer program, the program performing operations to determine common primitive features in the text segments, determine common composite features in the text segments; calculate a similarity measure based upon said primitive and composite features, and provide an output indicative of the similarity measure.

9. The system for determining similarity as defined by claim 8, wherein said primitive features are selected from the group including common single word, common noun phrase, synonyms, common semantic class of verbs, and common proper nouns.

10. The system for determining similarity as defined by claim 8, wherein said composite features are selected from the group including primitive feature order restrictions, primitive distance restrictions, and primitive type restrictions.

11. The system for determining similarity as defined by claim 8, wherein the processing operation of determining common primitive features includes:

identifying common primitive features;

assigning a value to said primitive features; and

normalizing said value.

12. The system for determining similarity as defined by claim 11, wherein the processing operation of normalizing includes at least one of normalizing for text segment length and normalizing for frequency of primitive occurrence.

13. The system for determining similarity as defined by claim 8, wherein said processing operation for determining common composite features includes:

identifying common primitive features;
assigning a value to said primitive features; and
normalizing said value.

14. The system for determining similarity as defined by claim 13, wherein said processing operation for normalizing includes at least one of normalizing for text segment length and normalizing for frequency of primitive occurrence.

15. The system for determining similarity as defined by claim 8, wherein the computer program includes a noun phrase identification subroutine, a synonym detection subroutine, a verb classifier subroutine and a word co-occurrence subroutine.

16. The system for determining similarity as defined by claim 8, further comprising a computer readable training corpus, and wherein the computer program includes a machine learning algorithm operatively coupled to the training corpus for learning and applying a rule set for determining similarity in small text segments.

FIG. 1

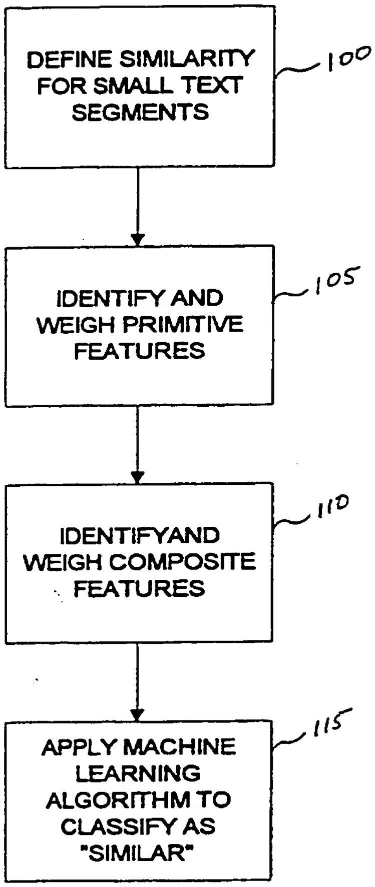

DEFINE SIMILARITY FOR SMALL TEXT SEGMENTS
100
IDENTIFY AND WEIGH PRIMITIVE FEATURES
105
IDENTIFYAND WEIGH COMPOSITE FEATURES
110
APPLY MACHINE LEARNING ALGORITHM TO CLASSIFY AS "SIMILAR"
115

FIG. 2

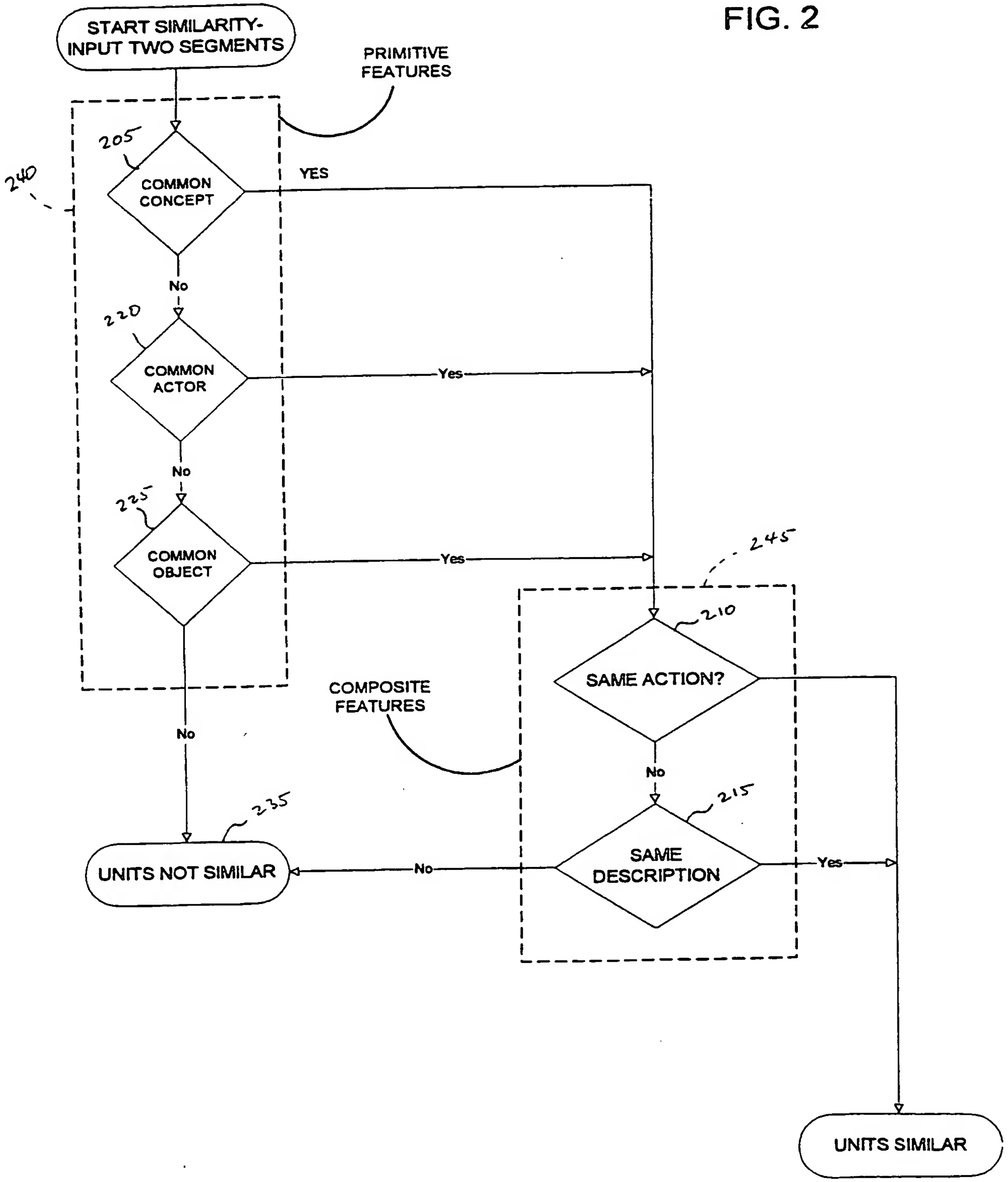

START SIMILARITY-
INPUT TWO SEGMENTS
PRIMITIVE FEATURES
205
240
COMMON CONCEPT
YES
No
220
COMMON ACTOR
Yes
No
225
COMMON OBJECT
Yes
245
210
SAME ACTION?
COMPOSITE FEATURES
No
No
215
235
UNITS NOT SIMILAR
No
SAME DESCRIPTION
Yes
UNITS SIMILAR

FIG. 3

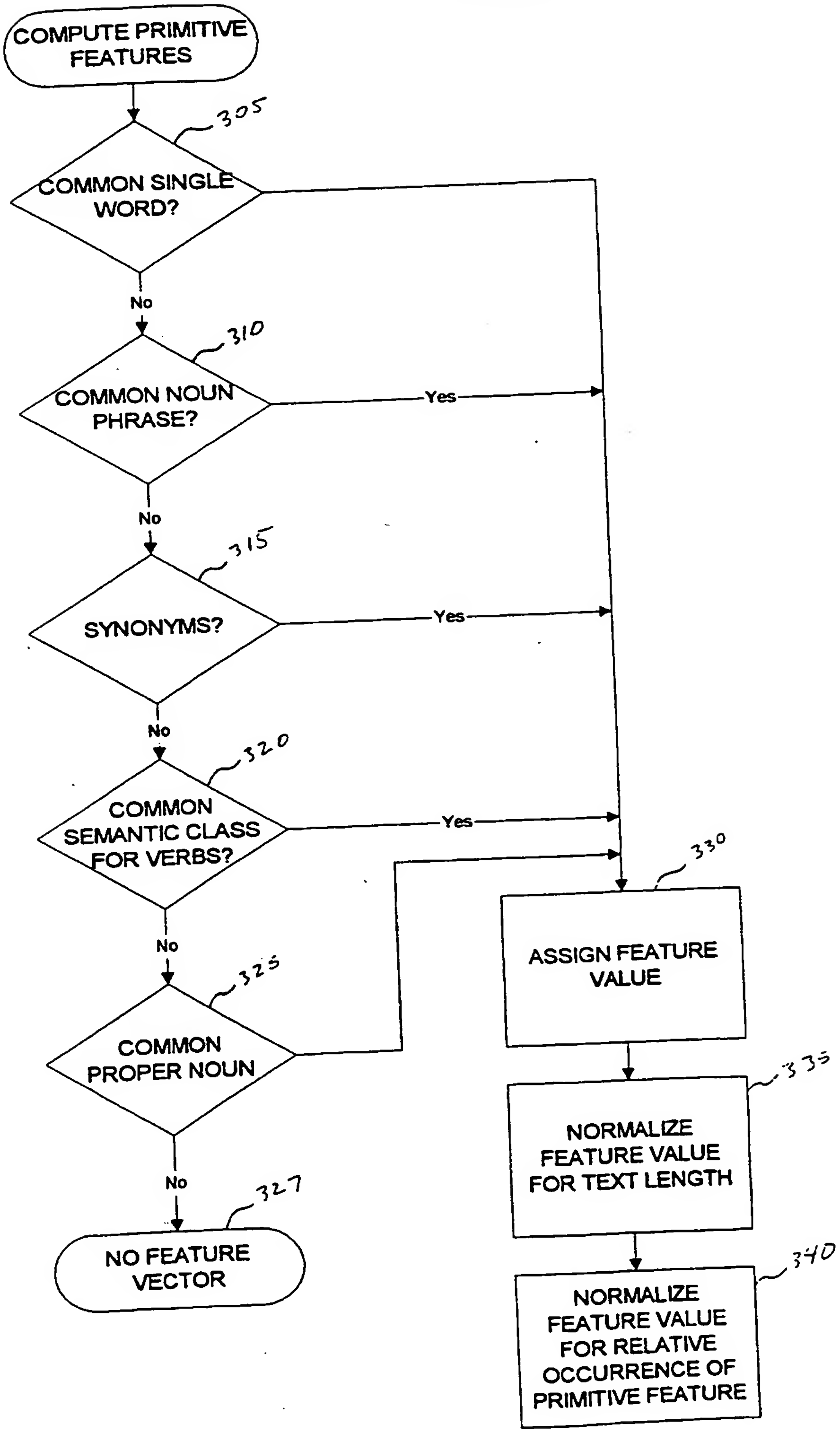

COMPUTE PRIMITIVE FEATURES
305
COMMON SINGLE WORD?
No
310
COMMON NOUN PHRASE?
Yes
No
315
SYNONYMS?
Yes
No
320
COMMON SEMANTIC CLASS FOR VERBS?
Yes
No
325
COMMON PROPER NOUN
No
327
NO FEATURE VECTOR
330
ASSIGN FEATURE VALUE
335
NORMALIZE FEATURE VALUE FOR TEXT LENGTH
340
NORMALIZE FEATURE VALUE FOR RELATIVE OCCURRENCE OF PRIMITIVE FEATURE

FIG. 4

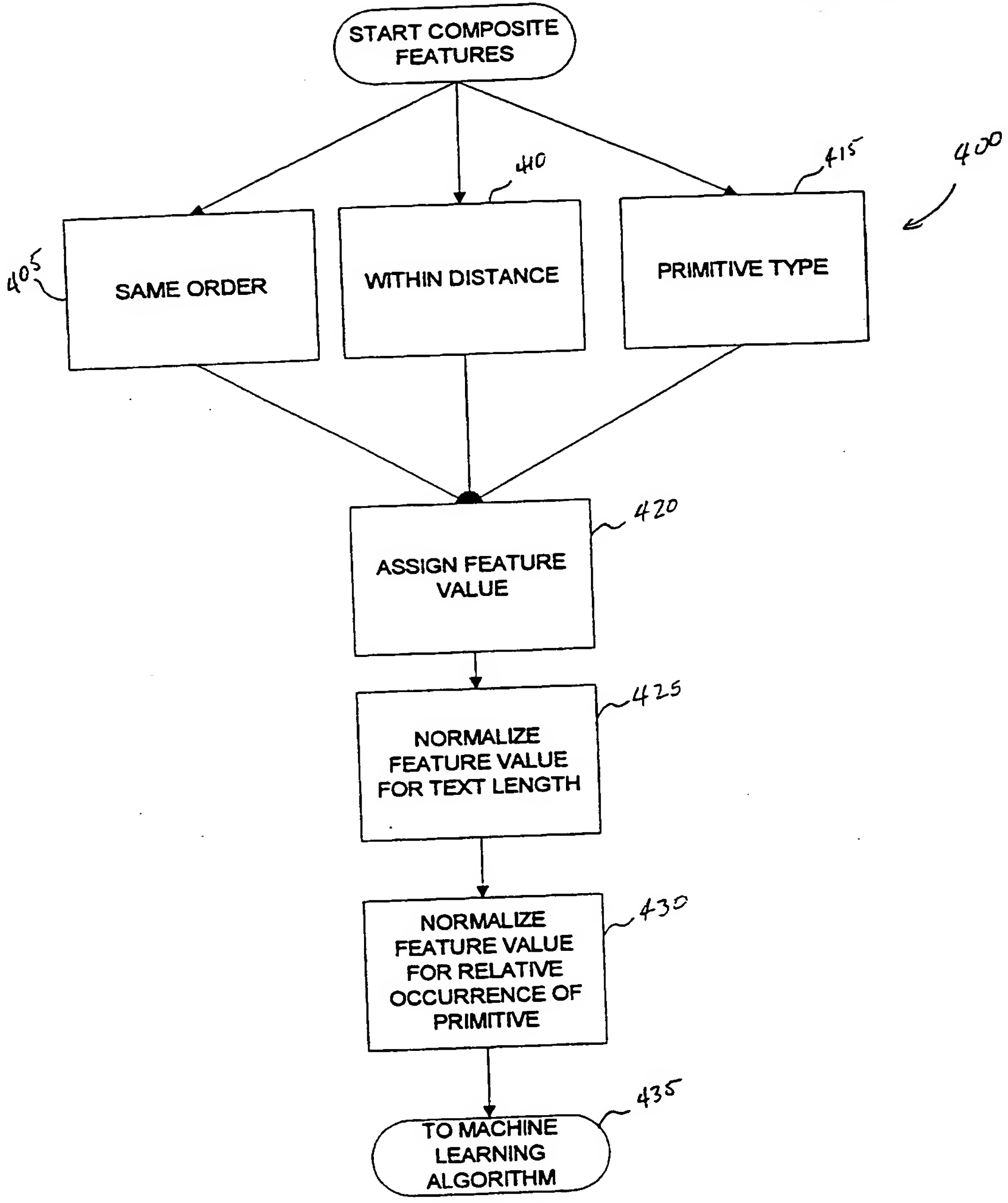

START COMPOSITE FEATURES
410
415
400
405
SAME ORDER
WITHIN DISTANCE
PRIMITIVE TYPE
420
ASSIGN FEATURE VALUE
425
NORMALIZE FEATURE VALUE FOR TEXT LENGTH
430
NORMALIZE FEATURE VALUE FOR RELATIVE OCCURRENCE OF PRIMITIVE
435
TO MACHINE LEARNING ALGORITHM

FIG. 5

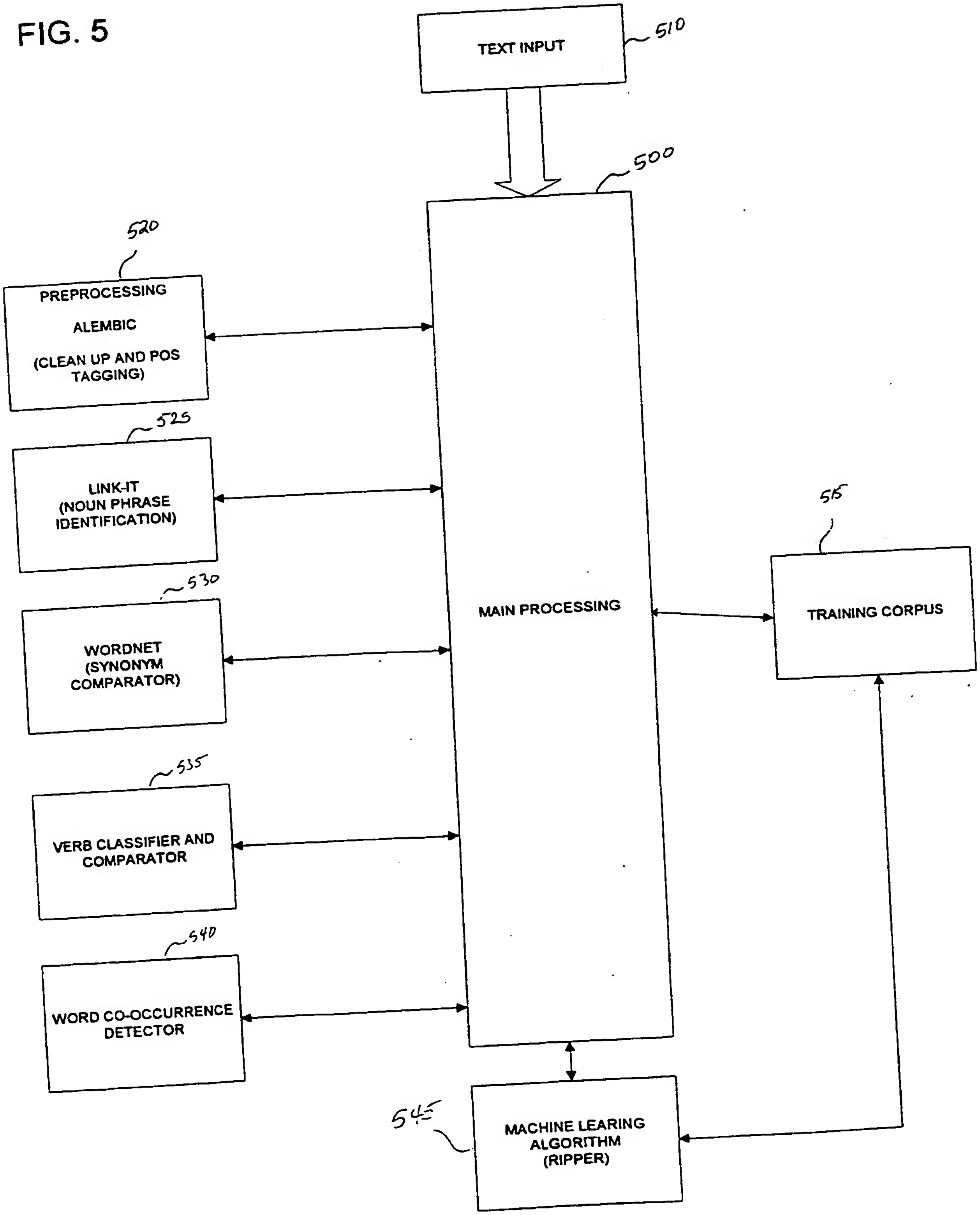

TEXT INPUT
510
500
520
PREPROCESSING
ALEMBIC
(CLEAN UP AND POS TAGGING)
525
LINK-IT
(NOUN PHRASE IDENTIFICATION)
530
WORDNET
(SYNONYM COMPARATOR)
535
VERB CLASSIFIER AND COMPARATOR
540
WORD CO-OCCURRENCE DETECTOR
MAIN PROCESSING
515
TRAINING CORPUS
545
MACHINE LEARING ALGORITHM
(RIPPER)

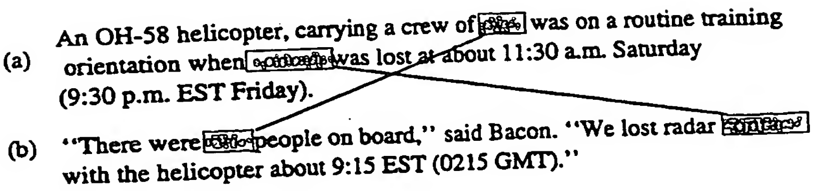
Fig 6 (a) An OH-58 helicopter, carrying a crew of two, was on a routine training orientation when contact was lost at about 11:30 a.m. Saturday (9:30 p.m. EST Friday).

Fig 6 (b) "There were two people on board," said Bacon. "We lost radar contact with the helicopter about 9:15 EST (0215 GMT)."

Fig 6 (c) An OH-58 U.S. military scout helicopter made an emergency landing in North Korea at about 9.15 p.m. EST Friday (0215 GMT Saturday), the Defense Department said.

Figure 1: Input text units (from the TDT pilot—corpus, topic 11).

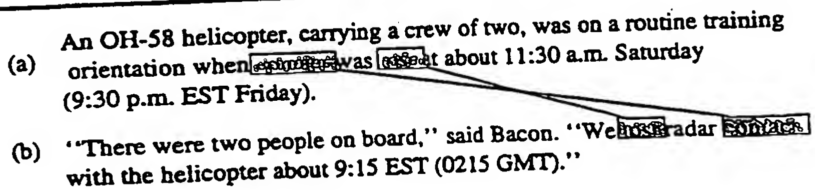
Fig 7



(a) An OH-58 helicopter, carrying a crew of [illegible] was on a routine training orientation when [illegible] was lost at about 11:30 a.m. Saturday (9:30 p.m. EST Friday).

(b) "There were [illegible] people on board," said Bacon. "We lost radar [illegible] with the helicopter about 9:15 EST (0215 GMT)."

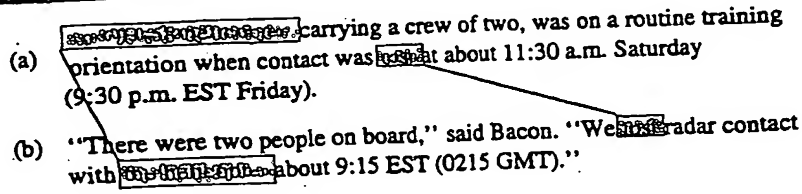
Fig 8



(a) An OH-58 helicopter, carrying a crew of two, was on a routine training orientation when [illegible] was [illegible]t about 11:30 a.m. Saturday (9:30 p.m. EST Friday).

(b) "There were two people on board," said Bacon. "We [illegible] radar [illegible] with the helicopter about 9:15 EST (0215 GMT)."

Fig 9



(a) [illegible] carrying a crew of two, was on a routine training orientation when contact was [illegible]t about 11:30 a.m. Saturday (9:30 p.m. EST Friday).

(b) "There were two people on board," said Bacon. "We [illegible] radar contact with [illegible] about 9:15 EST (0215 GMT)."

PATENT COOPERATION TREATY

From the INTERNATIONAL SEARCHING AUTHORITY

To: HENRY TANG
BAKER BOTTS LLP
30 ROCKEFELLER PLAZA
NEW YORK, NEW YORK 10112-0228

PCT

NOTIFICATION OF TRANSMITTAL OF THE INTERNATIONAL SEARCH REPORT OR THE DECLARATION

(PCT Rule 44.1)

Date of Mailing *(day/month/year)* 03 OCT 2000

Applicant's or agent's file reference
32550-PCT

FOR FURTHER ACTION See paragraphs 1 and 4 below

International application No.
PCT/US00/40238

International filing date *(day/month/year)*
19 JUNE 2000

Applicant
THE TRUSTEES OF COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

1. ☒ The applicant is hereby notified that the international search report has been established and is transmitted herewith.

Filing of amendments and statement under Article 19:
The applicant is entitled, if he so wishes, to amend the claims of the international application (see Rule 46):

When? The time limit for filing such amendments is normally 2 months from the date of transmittal of the international search report; however, for more details, see the notes on the accompanying sheet.

Where? Directly to the International Bureau of WIPO
34, chemin des Colombettes
1211 Geneva 20, Switzerland
Facsimile No.: (41-22) 740.14.35

For more detailed instructions, see the notes on the accompanying sheet.

2. □ The applicant is hereby notified that no international search report will be established and that the declaration under Article 17(2)(a) to that effect is transmitted herewith.

3. □ **With regard to the protest** against payment of (an) additional fee(s) under Rule 40.2, the applicant is notified that:

□ the protest together with the decision thereon has been transmitted to the International Bureau together with the applicant's request to forward the texts of both the protest and the decision thereon to the designated Offices.

□ no decision has been made yet on the protest; the applicant will be notified as soon as a decision is made.

4. **Further action(s):** The applicant is reminded of the following:

Shortly after **18 months** from the priority date, the international application will be published by the International Bureau. If the applicant wishes to avoid or postpone publication, a notice of withdrawal of the international application, or of the priority claim, must reach the International Bureau as provided in rules 90 *bis* 1 and 90 *bis* 3, respectively, before the completion of the technical preparations for international publication.

Within **19 months** from the priority date, a demand for international preliminary examination must be filed if the applicant wishes to postpone the entry into the national phase until 30 months from the priority date (in some Offices even later).

Within **20 months** from the priority date, the applicant must perform the prescribed acts for entry into the national phase before all designated Offices which have not been elected in the demand or in a later election within 19 months from the priority date or could not be elected because they are not bound by Chapter II.

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231
Facsimile No. (703) 305-3230

Authorized officer
JOSEPH THOMAS
James R. Matthews
Telephone No. (703) 308-3900

(See notes on accompanying sheet)

* Copy of Search Report ; References in Packet

PATENT COOPERATION TREATY

PCT

INTERNATIONAL SEARCH REPORT

(PCT Article 18 and Rules 43 and 44)

Applicant's or agent's file reference 32550-PCT	**FOR FURTHER ACTION** see Notification of Transmittal of International Search Report (Form PCT/ISA/220) as well as, where applicable, item 5 below.	
International application No. PCT/US00/40238	International filing date *(day/month/year)* 19 JUNE 2000	(Earliest) Priority Date *(day/month/year)* 18 JUNE 1999
Applicant THE TRUSTEES OF COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK		

This international search report has been prepared by this International Searching Authority and is transmitted to the applicant according to Article 18. A copy is being transmitted to the International Bureau.

This international search report consists of a total of 5 sheets.

☒ It is also accompanied by a copy of each prior art document cited in this report.

1. **Basis of the report**

a. With regard to the **language**, the international search was carried out on the basis of the international application in the language in which it was filed, unless otherwise indicated under this item.

□ the international search was carried out on the basis of a translation of the international application furnished to this Authority (Rule 23.1(b)).

b. With regard to any **nucleotide and/or amino acid sequence** disclosed in the international application, the international search was carried out on the basis of the sequence listing:

□ contained in the international application in written form.

□ filed together with the international application in computer readable form.

□ furnished subsequently to this Authority in written form.

□ furnished subsequently to this Authority in computer readable form.

□ the statement that the subsequently furnished written sequence listing does not go beyond the disclosure in the international application as filed has been furnished.

□ the statement that the information recorded in computer readable form is identical to the written sequence listing has been furnished.

2. □ **Certain claims were found unsearchable** (See Box I).

3. □ **Unity of invention is lacking** (See Box II).

4. With regard to the **title**,

☒ the text is approved as submitted by the applicant.

□ the text has been established by this Authority to read as follows:

5. With regard to the **abstract**,

□ the text is approved as submitted by the applicant.

☒ the text has been established, according to Rule 38.2(b), by this Authority as it appears in Box III. The applicant may, within one month from the date of mailing of this international search report, submit comments to this Authority.

6. The figure of the **drawings** to be published with the abstract is Figure No. 1

□ as suggested by the applicant.

☒ because the applicant failed to suggest a figure.

□ because this figure better characterizes the invention.

□ None of the figures.

International application No.
PCT/US00/40238

Box III TEXT OF THE ABSTRACT (Continuation of item 5 of the first sheet)

The technical features mentioned in the abstract do not include a reference sign between parentheses (PCT Rule 8.1(d)).

NEW ABSTRACT

A system and method are provided for determining similarity in short text segments. The method provides a definition of similarity which is appropriate for the small text setting (100). Small text segments are compared to determine if there exist common primitive features, such as words, noun phrases, synonyms, verbs with a common semantic class, proper nouns and the like (105). From the primitive features identified, the small text segments are evaluated to determine whether composite features are present (110). Composite features are defined as predetermined relationships between primitive features. The common primitive features and composite features are applied as inputs to an appropriate machine learning algorithm which is trained to ascertain a similarity measure based on the primitive and composite features common to the text segments (115).

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US00/40238

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) :G06F 17/21
US CL :704/10
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 704/1, 9, 10; 707/6, 531, 532

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Please See Extra Sheet.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 5,278,980 A (PEDERSEN et al) 11 January 1994, abstract; col. 1, line 13 to col. 6, line 18; and col. 7, line 55 to col. 16, line 22	1-16
Y	US 5,675,819 A (SCHUETZE) 07 October 1997, abstract; figs. 10-16; col. 1, line 6 to col. 5, line 15; and col. 13, line 40 to col. 21, line 21	1-16
Y	US 5,794,178 A (CAID et al) 11 August 1998, abstract; figs. 4-11 & 14-28; col. 1, line 21 to col. 3, line 42; col. 9, line 48 to col. 15, line 67; and col. 26, line 1 to col. 33, line 7	1-16

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search 05 SEPTEMBER 2000	Date of mailing of the international search report 03 OCT 2000
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer JOSEPH THOMAS James R. Matthews Telephone No. (703) 308-3900

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 5,893,095 A (JAIN et al) 06 April 1999, abstract; and col. 1, line 10 to col. 8, line 55	1-16
Y, P	US 5,943,669 A (NUMATA) 24 August 1999, abstract; col. 1, line 13 to col. 4, line 3; and col. 28, line 15 to col. 34, line 60	1-16

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US00/40238

B. FIELDS SEARCHED
Electronic data bases consulted (Name of data base and where practicable terms used):

EAST

search terms: similarity, normalization, text shorts/abstracts, semantic

32550 PCT

PATENT COOPERATION TREA.Y

From the INTERNATIONAL BUREAU

PCT

NOTIFICATION OF RECEIPT OF RECORD COPY

(PCT Rule 24.2(a))

To:

TANG, Henry
Baker Botts LLP
30 Rockefeller Plaza
New York, NY 10112-0228
ETATS-UNIS D'AMERIQUE

TO

Date of mailing (day/month/year) 11 September 2000 (11.09.00)	**IMPORTANT NOTIFICATION**
Applicant's or agent's file reference 32550-PCT	**International application No.** PCT/US00/40238

The applicant is hereby notified that the International Bureau has received the record copy of the international application as detailed below.

Name(s) of the applicant(s) and State(s) for which they are applicants:

THE TRUSTEES OF COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK (for all designated States except US)
KLAVANS, Judith, L. et al (for US)

International filing date	:	19 June 2000 (19.06.00)
Priority date(s) claimed	:	18 June 1999 (18.06.99)
Date of receipt of the record copy by the International Bureau	:	22 August 2000 (22.08.00)
List of designated Offices	:	

EP :AT,BE,CH,CY,DE,DK,ES,FI,FR,GB,GR,IE,IT,LU,MC,NL,PT,SE
National :JP,US

ATTENTION

The applicant should carefully check the data appearing in this Notification. In case of any discrepancy between these data and the indications in the international application, the applicant should immediately inform the International Bureau.

In addition, the applicant's attention is drawn to the information contained in the Annex, relating to:

☒ time limits for entry into the national phase

☒ confirmation of precautionary designations

☒ requirements regarding priority documents

A copy of this Notification is being sent to the receiving Office and to the International Searching Authority.

ON DOCKET FOR
2/18/01 Deadline

The International Bureau of WIPO **34, chemin des Colombettes** **1211 Geneva 20, Switzerland**	Authorized officer: R. Raissi
Facsimile No. (41-22) 740.14.35	Telephone No. (41-22) 338.83.38

ANNEX TO FORM PCT/IB/301

ernational application No.
PCT/US00/40238

INFORMATION ON TIME LIMITS FOR ENTERING THE NATIONAL PHASE

The applicant is reminded that the "national phase" must be entered before each of the designated Offices indicated in the Notification of Receipt of Record Copy (Form PCT/IB/301) by paying national fees and furnishing translations, as prescribed by the applicable national laws.

The time limit for performing these procedural acts is **20 MONTHS** from the priority date or, for those designated States which the applicant elects in a demand for international preliminary examination or in a later election, **30 MONTHS** from the priority date, provided that the election is made before the expiration of 19 months from the priority date. Some designated (or elected) Offices have fixed time limits which expire even later than 20 or 30 months from the priority date. In other Offices an extension of time or grace period, in some cases upon payment of an additional fee, is available.

In addition to these procedural acts, the applicant may also have to comply with other special requirements applicable in certain Offices. **It is the applicant's responsibility** to ensure that the necessary steps to enter the national phase are taken in a timely fashion. Most designated Offices do not issue reminders to applicants in connection with the entry into the national phase.

For detailed information about the procedural acts to be performed to enter the national phase before each designated Office, the applicable time limits and possible extensions of time or grace periods, and any other requirements, see the relevant Chapters of Volume II of the PCT Applicant's Guide. Information about the requirements for filing a demand for international preliminary examination is set out in Chapter IX of Volume I of the PCT Applicant's Guide.

GR and ES became bound by PCT Chapter II on 7 September 1996 and 6 September 1997, respectively, and may, therefore, be elected in a demand or a later election filed on or after 7 September 1996 and 6 September 1997, respectively, regardless of the filing date of the international application. (See second paragraph above.)

Note that only an applicant who is a national or resident of a PCT Contracting State which is bound by Chapter II has the right to file a demand for international preliminary examination.

CONFIRMATION OF PRECAUTIONARY DESIGNATIONS

This notification lists only specific designations made under Rule 4.9(a) in the request. It is important to check that these designations are correct. Errors in designations can be corrected where precautionary designations have been made under Rule 4.9(b). The applicant is hereby reminded that any precautionary designations may be confirmed according to Rule 4.9(c) before the expiration of 15 months from the priority date. If it is not confirmed, it will automatically be regarded as withdrawn by the applicant. There will be no reminder and no invitation. Confirmation of a designation consists of the filing of a notice specifying the designated State concerned (with an indication of the kind of protection or treatment desired) and the payment of the designation and confirmation fees. Confirmation must reach the receiving Office within the 15-month time limit.

REQUIREMENTS REGARDING PRIORITY DOCUMENTS

For applicants who have not yet complied with the requirements regarding priority documents, the following is recalled.

Where the priority of an earlier national, regional or international application is claimed, the applicant must submit a copy of the said earlier application, certified by the authority with which it was filed ("the priority document") to the receiving Office (which will transmit it to the International Bureau) or directly to the International Bureau, before the expiration of 16 months from the priority date, provided that any such priority document may still be submitted to the International Bureau before that date of international publication of the international application, in which case that document will be considered to have been received by the International Bureau on the last day of the 16-month time limit (Rule 17.1(a)).

Where the priority document is issued by the receiving Office, the applicant may, instead of submitting the priority document, request the receiving Office to prepare and transmit the priority document to the International Bureau. Such request must be made before the expiration of the 16-month time limit and may be subjected by the receiving Office to the payment of a fee (Rule 17.1(b)).

If the priority document concerned is not submitted to the International Bureau or if the request to the receiving Office to prepare and transmit the priority document has not been made (and the corresponding fee, if any, paid) within the applicable time limit indicated under the preceding paragraphs, any designated State may disregard the priority claim, provided that no designated Office may disregard the priority claim concerned before giving the applicant an opportunity to furnish the priority document within a time limit which is reasonable under the circumstances.

Where several priorities are claimed, the priority date to be considered for the purposes of computing the 16-month time limit is the filing date of the earliest application whose priority is claimed.

PCT/US00/40238

32850
PCA

PATENT COOPERATION TREATY

From the INTERNATIONAL BUREAU

PCT

NOTIFICATION CONCERNING SUBMISSION OR TRANSMITTAL OF PRIORITY DOCUMENT

(PCT Administrative Instructions, Section 411)

To:

TANG, Henry
Baker Botts LLP
30 Rockefeller Plaza
New York, NY 10112-0228
ETATS-UNIS D'AMERIQUE

Date of mailing (day/month/year) 17 October 2000 (17.10.00)	
Applicant's or agent's file reference 32550-PCT	**IMPORTANT NOTIFICATION**
International application No. PCT/US00/40238	International filing date (day/month/year) 19 June 2000 (19.06.00)
International publication date (day/month/year) Not yet published	Priority date (day/month/year) 18 June 1999 (18.06.99)

Applicant

THE TRUSTEES OF COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK et al

1. The applicant is hereby notified of the date of receipt (except where the letters "NR" appear in the right-hand column) by the International Bureau of the priority document(s) relating to the earlier application(s) indicated below. Unless otherwise indicated by an asterisk appearing next to a date of receipt, or by the letters "NR", in the right-hand column, the priority document concerned was submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b).

2. This updates and replaces any previously issued notification concerning submission or transmittal of priority documents.

3. An **asterisk(*)** appearing next to a date of receipt, in the right-hand column, denotes a priority document submitted or transmitted to the International Bureau but not in compliance with Rule 17.1(a) or (b). In such a case, **the attention of the applicant is directed** to Rule 17.1(c) which provides that no designated Office may disregard the priority claim concerned before giving the applicant an opportunity, upon entry into the national phase, to furnish the priority document within a time limit which is reasonable under the circumstances.

4. The **letters "NR"** appearing in the right-hand column denote a priority document which was not received by the International Bureau or which the applicant did not request the receiving Office to prepare and transmit to the International Bureau, as provided by Rule 17.1(a) or (b), respectively. In such a case, **the attention of the applicant is directed** to Rule 17.1(c) which provides that no designated Office may disregard the priority claim concerned before giving the applicant an opportunity, upon entry into the national phase, to furnish the priority document within a time limit which is reasonable under the circumstances.

Priority date	Priority application No.	Country or regional Office or PCT receiving Office	Date of receipt of priority document
18 June 1999 (18.06.99)	60/139,930	US	14 Sept 2000 (14.09.00)

The International Bureau of WIPO **34, chemin des Colombettes** **1211 Geneva 20, Switzerland** Facsimile No. (41-22) 740.14.35	Authorized officer Khemais BRAHMI Telephone No. (41-22) 338.83.38

WO 00/79426
PCT/US00/40238
32550

PATENT COOPERATION TREATY

From the INTERNATIONAL BUREAU PCT

To:
TANG, Henry
Baker Botts LLP
30 Rockefeller Plaza
New York, NY 10112-0228
ETATS-UNIS D'AMERIQUE

PCT

NOTICE INFORMING THE APPLICANT OF THE COMMUNICATION OF THE INTERNATIONAL APPLICATION TO THE DESIGNATED OFFICES

(PCT Rule 47.1(c), first sentence)

Date of mailing (day/month/year) 28 December 2000 (28.12.00)		
Applicant's or agent's file reference 32550-PCT	**IMPORTANT NOTICE**	
International application No. PCT/US00/40238	**International filing date** (day/month/year) 19 June 2000 (19.06.00)	**Priority date** (day/month/year) 18 June 1999 (18.06.99)
Applicant THE TRUSTEES OF COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK et al		

1. Notice is hereby given that the International Bureau has communicated, as provided in Article 20, the international application to the following designated Offices on the date indicated above as the date of mailing of this Notice:
   US

   In accordance with Rule 47.1(c), third sentence, those Offices will accept the present Notice as conclusive evidence that the communication of the international application has duly taken place on the date of mailing indicated above and no copy of the international application is required to be furnished by the applicant to the designated Office(s).

2. The following designated Offices have waived the requirement for such a communication at this time:
   EP,JP

   The communication will be made to those Offices only upon their request. Furthermore, those Offices do not require the applicant to furnish a copy of the international application (Rule 49.1(a-bis)).

3. Enclosed with this Notice is a ~~copy~~ of the international application as published by the International Bureau on 28 December 2000 (28.12.00) under No. WO 00/79426

REMINDER REGARDING CHAPTER II (Article 31(2)(a) and Rule 54.2)

If the applicant wishes to postpone entry into the national phase until 30 months (or later in some Offices) from the priority date, **a demand for international preliminary examination** must be filed with the competent International Preliminary Examining Authority before the expiration of 19 months from the priority date.

It is the applicant's sole responsibility to monitor the 19-month time limit.

Note that only an applicant who is a national or resident of a PCT Contracting State which is bound by Chapter II has the right to file a demand for international preliminary examination.

REMINDER REGARDING ENTRY INTO THE NATIONAL PHASE (Article 22 or 39(1))

If the applicant wishes to proceed with the international application in the **national phase**, he must, within 20 months or 30 months, or later in some Offices, perform the acts referred to therein before each designated or elected Office.

For further important information on the time limits and acts to be performed for entering the national phase, see the Annex to Form PCT/IB/301 (Notification of Receipt of Record Copy) and Volume II of the PCT Applicant's Guide.

2/18/01

The International Bureau of WIPO 34, chemin des Colombettes 1211 Geneva 20, Switzerland	Authorized officer J. Zahra
Facsimile No. (41-22) 740.14.35	Telephone No. (41-22) 338.83.38

Final

Encl. in pocket

PATENT COOPERATION TREATY

From the INTERNATIONAL BUREAU

PCT

NOTIFICATION OF ELECTION

(PCT Rule 61.2)

To:

Commissioner
US Department of Commerce
United States Patent and Trademark
Office, PCT
2011 South Clark Place Room
CP2/5C24
Arlington, VA 22202
ETATS-UNIS D'AMERIQUE
in its capacity as elected Office

Date of mailing (day/month/year) 27 November 2001 (27.11.01)	
International application No. PCT/US00/40238	**Applicant's or agent's file reference** 32550-PCT
International filing date (day/month/year) 19 June 2000 (19.06.00)	**Priority date** (day/month/year) 18 June 1999 (18.06.99)
Applicant KLAVANS, Judith, L. et al	

1. The designated Office is hereby notified of its election made:

☒ in the demand filed with the International Preliminary Examining Authority on:

03 January 2001 (03.01.01)

□ in a notice effecting later election filed with the International Bureau on:

2. The election ☒ was

□ was not

made before the expiration of 19 months from the priority date or, where Rule 32 applies, within the time limit under Rule 32.2(b).

The International Bureau of WIPO **34, chemin des Colombettes** **1211 Geneva 20, Switzerland** Facsimile No.: (41-22) 740.14.35	Authorized officer Imelda REHS Telephone No.: (41-22) 338.83.38

10/018108

TRANSMITTAL LETTER TO THE UNITED STATES RECEIVING OFFICE

Date	3 January 2001
International Application No.	PCT/US00/40238
Attorney Docket No.	32550-PCT

I. Certification under 37 CFR 1.10 (if applicable)

EK839852479US	3 January 2001
Express Mail mailing number	Date of Deposit

I hereby certify that the application/correspondence attached hereto is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and is addressed to Assistant Commissioner for Patents, Washington, D.C. 20231.

Leroy Chick	Leroy Chick
Signature of person mailing correspondence	Typed or printed name of person mailing correspondence

II. □ **New International Application**

TITLE		Earliest priority date (Day/Month/Year)

SCREENING DISCLOSURE INFORMATION: In order to assist in screening the accompanying international application for purposes of determining whether a license for foreign transmittal should and could be granted and for other purposes, the following information is supplied. (Note: check as many boxes as apply):

A. □ The invention disclosed was **not** made in the United States.

B. □ There is no prior U.S. application relating to this invention.

C. □ The following prior U.S. application(s) contain subject matter which is related to the invention disclosed in the attached international application. *(NOTE: priority to these applications may or may not be claimed on form PCT/RO/101 (Request) and this listing does not constitute a claim for priority).*

application no.		filed on	
application no.		filed on	

D. □ The present international application □ is identical □ contains less subject matter than that found in the prior U.S. application(s) identified in paragraph C.

E. □ The present international application □ contains additional subject matter not found in the prior U.S. application(s) identified in paragraph C. above. The additional subject matter is found on pages ______ and □ **DOES NOT ALTER** □ **MIGHT BE CONSIDERED TO ALTER** the general nature of the invention in a manner which would require the U.S. application to have been made available for inspection by the appropriate defense agencies under 35 U.S.C. 181 and 37 CFR 5.1. See 37 CFR 5.15

III. □ **A Response to an Invitation from the RO/US.** The following document(s) is (are) enclosed:

A. □ A Request for An Extension of Time to File a Response

B. □ A Power of Attorney (General or Regular)

C. □ Replacement pages:

pages		of the request (PCT/RO/101)	pages		of the figures
pages		of the description	pages		of the abstract
pages		of the claims			

D. □ Submission of Priority Documents

Priority document		Priority document	

E. □ Fees as specified on attached Fee Calculation sheet form PCT/RO/101 annex

IV. □ **A Request for Rectification under PCT 91** □ **A Petition** □ **A Sequence Listing Diskette**

V. ☒ **Other (please specify):** Demand for International Preliminary Examination (4 sheets), Fee Calculation Sheet, a postcard and a check in the amount of $627.

The person signing this form is the:		
	□ Applicant	Paul D. Ackerman Typed name of signer
	☒ Attorney/Agent (Reg. No.) 39,891	
	□ Common Representative	Signature

The demand must be filed directly [illegible] competent International Preliminary Examining Authority or, if two or more Authorities are with the one chosen by the applicant. The full name or two-letter code of that Authority may be indicated by the applicant on the line

IPEA/ US

PCT

CHAPTER II

DEMAND

under Article 31 of the Patent Cooperation Treaty:
The undersigned requests that the international application specified below be the subject of international preliminary examination according to the Patent Cooperation Treaty and hereby elects all eligible States (except where otherwise indicated).

For International Preliminary Examining Authority use only

Identification of IPEA	Date of receipt of DEMAND

Box No. I IDENTIFICATION OF THE INTERNATIONAL APPLICATION		Applicant's or agent's file reference 32550-PCT
International application No. PCT/US00/40238	International filing date *(day/month/year)* 19 June 2000 (19.06.00)	(Earliest) Priority date *(day/month/year)* 18 June 1999 (18.06.99)

Title of invention
SYSTEM AND METHOD FOR DETECTING TEXT SIMILARITY OVER SHORT PASSAGES

Box No. II APPLICANT(S)

Name and address: *(Family name followed by given name; for a legal entity, full official designation. The address must include postal code and name of country.)* THE TRUSTEES OF COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK 116th Street and Broadway New York, NY 10027 US	Telephone No.: Facsimile No.: Teleprinter No.:
State *(that is, country)* of nationality: US	State *(that is, country)* of residence: US

Name and address: *(Family name followed by given name; for a legal entity, full official designation. The address must include postal code and name of country.)*

KLAVANS, JUDITH L.
40 South Drive
Hasting-on-Hudson, NY 10706
US

State *(that is, country)* of nationality: US	State *(that is, country)* of residence: US

Name and address: *(Family name followed by given name; for a legal entity, full official designation. The address must include postal code and name of country.)*

ESKIN, ELEAZAR
935 Stanford Street
Santa Monica, CA 90403
US

State *(that is, country)* of nationality: US	State *(that is, country)* of residence: US

☒ Further applicants are indicated on a continuation sheet.

Sheet No. .2.

International application No.
PCT/US00/40238

Continuation of Box No. II APPLICANT(S)

If none of the following sub-boxes is used, this sheet is not to be included in the demand.

Name and address: *(Family name followed by given name; for a legal entity, full official designation. The address must include postal code and name of country.)*

HATZIVASSILOGLOU, VASILEIOS
452 Riverside Drive, Apt. 41
New York, NY 10027
US

State *(that is, country)* of nationality:
GR

State *(that is, country)* of residence:
US

Name and address: *(Family name followed by given name; for a legal entity, full official designation. The address must include postal code and name of country.)*

State *(that is, country)* of nationality:

State *(that is, country)* of residence:

Name and address: *(Family name followed by given name; for a legal entity, full official designation. The address must include postal code and name of country.)*

State *(that is, country)* of nationality:

State *(that is, country)* of residence:

Name and address: *(Family name followed by given name; for a legal entity, full official designation. The address must include postal code and name of country.)*

State *(that is, country)* of nationality:

State *(that is, country)* of residence:

□ Further applicants are indicated on another continuation sheet.

Sheet No. .3.

International application No.
HATZIVASSILOGLOU,

Box No. III AGENT OR COMMON REPRESENTATIVE; OR ADDRESS FOR CORRESPONDENCE

The following person is ☒ agent □ common representative

and ☒ has been appointed earlier and represents the applicant(s) also for international preliminary examination.

□ is hereby appointed and any earlier appointment of (an) agent(s) /common representative is hereby revoked.

□ is hereby appointed, specifically for the procedure before the International Preliminary Examining Authority. in addition to the agent(s)/common representative appointed earlier.

Name and address: *(Family name followed by given name; for a legal entity, full official The address must include postal code and name of country.)* TANG, HENRY and ACKERMAN, PAUL D. Baker Botts LLP 30 Rockefeller Plaza New York, NY 10112 US	Telephone No.: (212) 705-5000 Facsimile No.: (212) 705-5020 Teleprinter No.:

□ Address for correspondence: Mark this check-box where no agent or common representative is/has been appointed and the space above is used instead to indicate a special address to which correspondence should be sent.

Box No. IV BASIS FOR INTERNATIONAL PRELIMINARY EXAMINATION

Statement concerning amendments:*

1. The applicant wishes the international preliminary examination **to start on the basis of:**

☒ the international application as originally filed.

the description □ as originally filed
□ as amended under Article 34

the claims □ as originally filed
□ as amended under Article 19 (together with any accompanying statement)
□ as amended under Article 34

the drawings □ as originally filed
□ as amended under Article 34

2. □ The applicant wishes any amendment to the claims under Article 19 to be considered as reversed.

3. □ The applicant wishes the start of the international preliminary examination **to be postponed** until the expiration of 20 months from the priority date unless the International Preliminary Examing Authority receives a copy of any amendments made under Article 19 or a notice from the applicant that he does not wish to make such amendments (Rule 69.1(d)). *(This check-box may be marked only where the time limit under Article 19 has not yet expired.)*

* Where no check-box is marked, international preliminary examination will start on the basis of the international application as originally filed or, where a copy of amendments to the claims under Article 19 and/or amendments of the international application under Article 34 are received by the International Preliminary Examining Authority before it has begun to draw up a written opinion or the international preliminary examination report, as so amended.

Language for the purposes of international preliminary examination: English

☒ which is the language in which the international application was filed.

□ which is the language of a translation furnished for the purposes of international search.

□ which is the language of publication of the international application.

□ which is the language of the translation (to be) furnished for the purposes of international preliminary examination.

Box No. V ELECTION OF STATES

The applicant hereby **elects all eligible States** *(that is, all States which have been designated and which are bound by Chapter II of the PCT)*

excluding the following States which the applicant wishes **not to elect:**

Sheet No. 3.

International application No.
HATZIVASSILOGLOU,

Box No. VI CHECK LIST

The demand is accompanied by the following elements, in the language referred to in Box No. IV, for the purposes of international preliminary examination:

			For International Preliminary Examining Authority use only received	 not received
1.	translation of international application	: sheets	□	□
2.	amendments under Article 34	: sheets	□	□
3.	copy (or where required, translation) of amendments under Article 19	: sheets	□	□
4.	copy (or, where required, translation) of statement under Article 19	: sheets	□	□
5.	letter	: sheets	□	□
6.	other *(specify)*	: sheets	□	□

The demand is also accompanied by the item(s) marked below:

1. ☒ fee calculation sheet
2. □ separate signed power of attorney
3. □ copy of general power of attorney; reference number, if any:
4. □ statement explaining lack of signature
5. □ nucleotide and or amino acid sequence listing in computer readable form
6. ☒ other *(specify)*: Transmittal Letter

Box No. VII SIGNATURE OF APPLICANT, AGENT OR COMMON REPRESENTATIVE

Next to each signature, indicate the name of the person signing and the capacity in which the person signs (if such capacity is no obvious from reading the demand).

Paul D. Ackerman (Agent)

For International Preliminary Examining Authority use only

1. Date of actual receipt of DEMAND:

2. Adjusted date of receipt of demand due to CORRECTIONS under Rule 60.1(b):

3. □ The date of receipt of the demand is AFTER the expiration of 19 months from the priority date and item 4 or 5, below, does not apply. □ The applicant has been informed accordingly.

4. □ The date of receipt of the demand is WITHIN the period of 19 months from the priority date as extended by virtue of Rule 80.5.

5. □ Although the date of receipt of the demand is after the expiration of 19 months from the priority date, the delay in arrival i EXCUSED pursuant to Rule 82.

For International Bureau use only

Demand received from IPEA on:

CHAPTER II

PCT

FEE CALCULATION SHEET

Annex to the Demand for international preliminary examination

International application No.	PCT/US00/40238
Applicant's or agent's file reference	32550-PCT

For International Preliminary Examining Authority use only

Date stamp of the IPEA

Applicant
THE TRUSTEES OF COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

Calculation of prescribed fees

1. Preliminary examination fee	490.00	P
2. Handling fee *(Applicants from certain States are entitled to a reduction of 75% of the handling fee. Where the applicant is (or all applicants are) so entitled, the amount to be entered at H is 25% of the handling fee.)*	137.00	H
3. Total of prescribed fees Add the amounts entered at P and H and enter total in the TOTAL box	627.00 TOTAL	

Mode of Payment

□ authorization to charge deposit account with the IPEA (see below)

☒ cheque

□ postal money order

□ bank draft

□ cash

□ revenue stamps

□ coupons

□ other *(specify)*:

Deposit Account Authorization *(this mode of payment may not be available at all IPEAs)*

The IPEA/ US □ is hereby authorized to charge the total fees indicated above to my deposit account.

☒ *(this check-box may be marked only if the conditions for deposit accounts of the IPEA so permit)* is hereby authorized to charge any deficiency or credit any overpayment in the total fees indicated above to my deposit account.

02-4377	3 January 2000	
Deposit Account Number	Date *(day/month/year)*	Signature

See Notes to the fee calculation sheet

~~FR~~

~~PB~~A

PATENT COOPERATION TREATY

From the
INTERNATIONAL PRELIMINARY EXAMINING AUTHORITY

To:
HENRY TANG
BAKER BOTTS LLP
30 ROCKEFELLER PLAZA
NEW YORK, NY 10112 0228

PCT

NOTIFICATION OF RECEIPT OF DEMAND BY COMPETENT INTERNATIONAL PRELIMINARY EXAMINING AUTHORITY

(PCT Rules 59.3(e) and 61.1(b), first sentence and Administrative Instructions, Section 601(a))

Date of mailing *(day/month/year)* 25 SEP 2001	
Applicant's or agent's file reference 32550-PCT	**IMPORTANT NOTIFICATION**

International application No. PCT/US00/40238	International filing date *(day/month/year)* 19 JUN 00	Priority date *(day/month/year)* 18 JUN 99

Applicant
THE TRUSTEES OF COLUMBIA UNIVERSITY IN THE CITY OF

1. The applicant is hereby **notified** that this International Preliminary Examining Authority considers the following date as the date of receipt of the demand for international preliminary examination of the international application:

   03 Jan 2001

2. That date of receipt is:

   ☑ the actual date of receipt of the demand by this Authority (Rule 61.1(b)).

   □ the actual date of receipt of the demand on behalf of this Authority (Rule 59.3(e)).

   □ the date on which this Authority has, in response to the invitation to correct defects in the demand (Form PCT/IPEA/404), received the required corrections.

3. □ **ATTENTION:** That date of receipt is **AFTER** the expiration of 19 months from the priority date. Consequently, the election(s) made in the demand does (do) not have the effect of postponing the entry into the national phase until 30 months from the priority date (or later in some Offices) (Article 39(1)). Therefore, the acts for entry into the national phase must be performed within 20 months from the priority date (or later in some Offices) (Article 22). For details, see the *PCT Applicant's Guide*, Volume II.

   □ *(If applicable)* This notification confirms the information given by telephone, facsimile transmission or in person on:

4. Only where paragraph 3 applies, a copy of this notification has been sent to the International Bureau.

Name and mailing address of the IPEA/ Assistant Commissioner for Patent Box PCT Washington, D.C. 20231 Attn:RO/US Facsimile No. 703-305-3230	Authorized officer D. Russell for Seg Hosford Telephone No. 703 305 3680

10/018,908

PATENT COOPERATION TREATY

PCT

INTERNATIONAL PRELIMINARY EXAMINATION REPORT

(PCT Article 36 and Rule 70)

Applicant's or agent's file reference 32550-PCT	**FOR FURTHER ACTION**	See Notification of Transmittal of International Preliminary Examination Report (Form PCT/IPEA/416)
International application No. PCT/US00/40238	International filing date *(day/month/year)* 19 June 2000 (19.06.2000)	Priority date *(day/month/year)* 18 June 1999 (18.06.1999)
International Patent Classification (IPC) or national classification and IPC IPC(7): G06F 17/21 and US Cl.: 704/10, 1, 9; 707/6,531, 532		
Applicant THE TRUSTEES OF COLUMBIA UNIVERSITY IN THE CITY OF		

1. This international preliminary examination report has been prepared by this International Preliminary Examining Authority and is transmitted to the applicant according to Article 36.

2. This REPORT consists of a total of 3 sheets, including this cover sheet.

   □ This report is also accompanied by ANNEXES, i.e., sheets of the description, claims and/or drawings which have been amended and are the basis for this report and/or sheets containing rectifications made before this Authority (see Rule 70.16 and Section 607 of the Administrative Instructions under the PCT).

   These annexes consist of a total of 0 sheets.

3. This report contains indications relating to the following items:

   I ☒ Basis of the report
   II □ Priority
   III □ Non-establishment of report with regard to novelty, inventive step and industrial applicability
   IV □ Lack of unity of invention
   V ☒ Reasoned statement under Article 35(2) with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement
   VI □ Certain documents cited
   VII □ Certain defects in the international application
   VIII □ Certain observations on the international application

Date of submission of the demand 03 January 2001 (03.01.2001)	Date of completion of this report 08 September 2002 (08.09.2002)
Name and mailing address of the IPEA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703)305-3230	Authorized officer Marsha D. Banks-Harold Telephone No. 703 3053900

INTERNATIONAL PRELIMINARY EXAMINATION REPORT

International application No.
PCT/US00/40238

I. Basis of the report

1. With regard to the elements of the international application:*

☒ the international application as originally filed.

☒ the description:
pages 1-12 as originally filed
pages NONE, filed with the demand
pages NONE, filed with the letter of ______.

☒ the claims:
pages 13-15, as originally filed
pages NONE, as amended (together with any statement) under Article 19
pages NONE, filed with the demand
pages NONE, filed with the letter of ______.

☒ the drawings:
pages 1-8, as originally filed
pages NONE, filed with the demand
pages NONE, filed with the letter of ______.

□ the sequence listing part of the description:
pages NONE, as originally filed
pages NONE, filed with the demand
pages NONE, filed with the letter of ______.

2. With regard to the language, all the elements marked above were available or furnished to this Authority in the language in which the international application was filed, unless otherwise indicated under this item.
These elements were available or furnished to this Authority in the following language ____ which is:

□ the language of a translation furnished for the purposes of international search (under Rule23.1(b)).

□ the language of publication of the international application (under Rule 48.3(b)).

□ the language of the translation furnished for the purposes of international preliminary examination(under Rules 55.2 and/or 55.3).

3. With regard to any **nucleotide and/or amino acid sequence** disclosed in the international application, the international preliminary examination was carried out on the basis of the sequence listing:

□ contained in the international application in printed form.

□ filed together with the international application in computer readable form.

□ furnished subsequently to this Authority in written form.

□ furnished subsequently to this Authority in computer readable form.

□ The statement that the subsequently furnished written sequence listing does not go beyond the disclosure in the international application as filed has been furnished.

□ The statement that the information recorded in computer readable form is identical to the written sequence listing has been furnished.

4. ☒ The amendments have resulted in the cancellation of:

☒ the description, pages NONE

☒ the claims, Nos. NONE

☒ the drawings, sheets/~~fig~~ NONE

5. □ This report has been established as if (some of) the amendments had not been made, since they have been considered to go beyond the disclosure as filed, as indicated in the Supplemental Box (Rule 70.2(c)).**

* *Replacement sheets which have been furnished to the receiving Office in response to an invitation under Article 14 are referred to in this report as "originally filed" and are not annexed to this report since they do not contain amendments (Rules 70.16 and 70.17).*
*** Any replacement sheet containing such amendments must be referred to under item 1 and annexed to this report.*

V. Reasoned statement under Rule 66.2(a)(ii) with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement

1. STATEMENT

Novelty (N)	Claims	4-7 and 11-14	YES
	Claims	1-3,8-10 and 15-16	NO
Inventive Step (IS)	Claims	NONE	YES
	Claims	1-16	NO
Industrial Applicability (IA)	Claims	1-16	YES
	Claims	NONE	NO

2. CITATIONS AND EXPLANATIONS

1. Claims 1-3, 8-10, and 15-16 lack novelty under PCT Article 33(2) as being anticipated by Kupiec (US 5,696,962 A).
(A) As per claim 1, Kupiec teaches a method for computerized information retrieval using shallow linguistic analysis, and for determining similarity in text segments (Kupiec; col. 33, lines 12-23), comprising the steps of:
"determining features such as part-of-speech information, noun phrase, verbs, synonyms, and hyponyms(reads on "primitive features") (Kupiec; col. 7, line 63 to col. 10, line 8, and col. 12, lines 51-65);
determining features such as proximity, order, and constraints (reads on "composite features") (Kupiec; col. 6, lines 45-57; and col. 33, line 25 to col. 34, line 55); and
matching like phrases by calculating similarity measures therefrom (Kupiec; col. 3, line 44-45; col. 32, line 64 to col. 33, line 23; and col. 35, line 5 to col. 36, line 43).,

(B) As per claim 2, note col. 7, line 63 to col. 10, line 8 and col. 12, lines 51-65 of Kupiec.
(C) As per claim 3, note col. 6, lines 45-57 and col. 33, line 25 to col. 34, line 55 of Kupiec.
(D) Claims 8-10 differ from claims 1-3 by reciting system elements such as an interface circuit and a main processing section operating under the control of a computer program. As per these limitations, Kupiec system has a user interface (7) and runs on a programmed CPU (5) and memory (6) (Kupiec; fig.. 1 and col. 5, lines 42-58). The remaining limitations of claims 8-10 are as addressed above in tire discussion of claims 1-3, and incorporated herein.
(E) As per claims 15-16, Kupiec discloses the use of part-of-speech taggers and phrase recognizers, as well as the training of text via a Hidden Markov Model (HMM) estimations (reads on "machine learning algorithm") (Kupiec; col. 8, lines 46 to col. 10, line 8; and col. 39, line 64 to col. 40, line 10).
11. Claims 4-7 and 11-14 lack an inventive step under PCT Article 33(3) as being obvious over Kupiec (US 5,696,962 A) in view of Schuetze (US 5,675,819 A).
(A) As per claims 4-5, Kupiec discloses the use of different ranking or prioritization criteria based on the frequency of some words within retrieved documents or the text corpus as a whole (Kupiec; col. 26, lines 22-28), but fails to expressly teach the normalizing of primitive features leaving assigned values and according to text segment length or frequency of word occurrence. However, this is known in the art, as evidenced by Schuetze.
In particular, Schuetze discloses computing context vectors for word (reads on "assigning values"), and then normalizing the context vectors (Schuetze; col. 17, lines 56 to col. 18, line 10 and fig. 10).
One having ordinary skill in the art at the tune of the invention would have found it obvious to assign values to the query features (such as part-of-speech information, noun phrase, verbs, synonyms, and hyponyms which read on "primitive features") and to normalizing these values with the motivation of improving retrieval performance for non-literal matches with queries (Schuetze; col. 4, lines 13-15).
(B) Claims 6-7, 11-12, and 13-14 repeat tire same limitations of claims 4-5 are therefore obvious for the same reasons given above for claims 4-5.

------------ NEW CITATIONS ------------
US 5,696,962 A (KUPIEC) 09 December 1997, see abstract; fig. 1, col. 3, lines 44-45; col. 5, lines 42-58; col. 6, line 45-57; col. 7, line 63 to col. 10, line 8; col. 12, lines 51-65; col. 26, lines 22-28; col. 32, line 64 to col. 34, line 55; col. 35, line 5 to col. 36, line 43; and col. 39, line 64 to col. 40, line 10.